ARI Research Note 90-33

Creating Algorithms as an Aid to Judgment

Sarah Lichtenstein, Donald MacGregor,
and Paul Slovic

Perceptronics, Inc.

for

Contracting Officer's Representative
Michael Drillings

Basic Research
Michael Kaplan, Director

June 1990

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

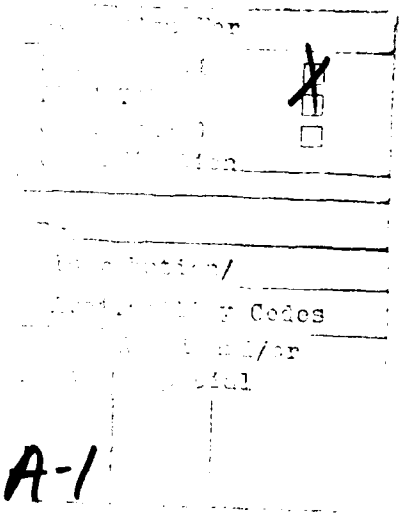
JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for
the Department of the Army

Perceptronics, Inc.

Technical review by

Nehama Babin



NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) PDIR-1129-8-87	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 90-33
6a. NAME OF PERFORMING ORGANIZATION Perceptronics, Inc. 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code) 6271 Variel Avenue Woodland Hills, CA 91367	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI-BR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-83-C-0198
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102B; PROJECT NO. 74F; TASK NO; WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) Creating Algorithms as an Aid to Judgment	
12. PERSONAL AUTHOR(S) Lichtenstein, Saran; MacGregor, Donald; and Slovic, Paul (Perceptronics, Inc.)	
13a. TYPE OF REPORT Interim	13b. TIME COVERED FROM 83/04 TO 87/08
14. DATE OF REPORT (Year, Month, Day) 1990, June	15. PAGE COUNT 38
16. SUPPLEMENTARY NOTATION Contracting Officer's Representative, Michael Drillings	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Decision aiding
Algorithmic decomposition
Misinformation

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The strategy for aiding judgment presented in this report is algorithmic decomposition. To use this approach, a complicated or unknown quantity is decomposed into a number of sub-problems that are more manageable or can be estimated more readily. Answers to the component parts of the problem are then combined according to a set of rules (an algorithm) to yield an answer to the original problem. In this experiment we gave subjects questions concerning facts they would be unlikely to know but could estimate and provided subjects with a wrong answer. The subjects' task was to decide whether the given wrong answer was too high or too low. After completing four such items, subjects were given tutorials on how to create algorithms, based on facts they knew or could estimate, to help them in their task. They then completed four more items under instructions to write an algorithm for each one. These efforts to teach subjects to create their own algorithms were successful in the sense that most subjects were able to write algorithms for the questions we gave them. However, the

(Continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Michael Drillings	22b. TELEPHONE (Include Area Code) (202) 274-8722 22c. OFFICE SYMBOL BR

DD Form 1473, JUN 86 *Previous editions are obsolete.*

ARI Research Note 90-33

19. ABSTRACT (Continued)

increase in accuracy of their decisions as a result of creating and using algorithms, from 65% to 69%, was only marginally significant and unimpressive in size.

ACKNOWLEDGMENT

We are grateful to Anna Gilson Weathers, who was one of the coders.

CREATING ALGORITHMS AS AN AID TO JUDGMENT

A vital task that people often undertake is to provide answers to important questions of fact. For example, experts make judgments about the likely value of a numerical quantity (e.g., "How many enemy troops are deployed in that area?" "How long will it take to repair the system?") or the chances that a particular goal will be achieved (e.g., "What is the probability that the system will be operational by tomorrow?"). If the facts are readily at hand and known for certain, responding to such queries might be a simple matter. If the facts aren't available, then the best one can do is formulate a judgment about the issue at hand based on whatever information is available. When time and resources are relatively limited, this may mean constructing a response by relying on relevant knowledge one has already accumulated in memory.

How can the contents of memory best be accessed to take the most complete advantage of what one knows relevant to a point of fact? The strategy here presented for aiding judgment is algorithmic decomposition. The essence of this approach is that a complicated or unknown quantity is decomposed into a number of sub-problems that are more manageable or can be estimated more readily. Answers to the component parts of the problem are then combined according to a set of rules (an algorithm) to yield an answer to the original problem.

Decomposition strategies have diverse application and play an important role in the operation of computers (e.g., Goodman & Hedetniemi, 1977) and decision-making systems (e.g., Raiffa,

1968). However, the pitfalls associated with employing algorithmic decomposition as a strategy for aiding people in organizing their knowledge about a problem have received little attention.

In previous research (MacGregor, Lichtenstein & Slovic, in press), we showed that when people are given algorithms written by the experimenters, their accuracy in estimating unknown quantities is better than performance without the algorithms. This result is encouraging, but it leaves a crucial step untested: Can people be trained to create their own algorithms, without receiving any training about the uncertain quantities? Will instructions about how to estimate the height of the Empire State Building from one's estimates of the average height per floor and the number of floors teach people how to construct their own algorithms for, for example, the number of pounds of potato chips consumed yearly in this country? If such training is successful, will accuracy be improved? The present paper is addressed to these questions.

In addition, we have a continuing interest in the effect of decision aids on confidence. In assessing one's confidence, the amount of knowledge one has available should serve as an important guide; the less one knows, the less confident one should be. Strategies that assist in the retrieval of knowledge can be expected to exert some influence on the confidence people attach to their judgments. The danger here is that a decision aid might, because it appears to be a help, increase confidence

without any corresponding increase in accuracy.

For several years we have tried unsuccessfully to teach our subjects how to create their own algorithms. The approach taken here differs from our earlier efforts in two respects. First, we have chosen uncertain quantities for which we were able to think of very simple algorithms, involving, in most cases, only two or three steps. Secondly, instead of asking the subject to produce an answer, we ask them to evaluate the answer provided for each question by the experimenters (which, we assure the subjects, is the wrong answer). Of course, building a successful algorithm requires coming up with an answer. But for this task the subjects do not have to worry about the accuracy of their own answers. They use these answers, instead, to decide whether the answer we have given them is wrong because it is too high or because it is too low.

Thus the overall plan of the present experiment was to present subjects with four questions and answers, asking them whether our answer is too high or too low. Following these control questions, we gave subjects a tutorial in how to write algorithms. After the tutorial, the subjects received four new questions and answers. The task was the same, but this time they were asked to write and use algorithms.

Receiving both a question and an answer, with the task of evaluating the answer, is a common experience. For example, questionable statistics abound in the news media (Singer, 1971). Can three million dollars really be the value of those

confiscated drugs? Is the U.S. really fourteenth in the world in child mortality rates?

Providing an answer also makes possible two different approaches to building algorithms. One can forget the answer, temporarily, and find one's own answer by constructing an algorithm. One then compares the new answer with the one provided. We call this the Forward approach. Alternatively, one can use the answer provided, decomposing it via an algorithm until one arrives at an estimate of some other quantity. This new estimate, which is an implication of the answer given, can then be judged directly. We call this the Backward approach. In the present experiment we test the effectiveness of both approaches.

Design

Table 1 shows the eight questions used in the experiment, along with the correct answer (found in reference books), the factor used to arrive at the answers given to subjects, and the high and low answers given to subjects.

Insert Table 1 about here

The questions were chosen to suggest relatively simple algorithms, for example, "Oregon is approximately in the shape of a rectangle 200 miles from North to South and 500 miles from East to West. Therefore the area is about 200 x 500 = 100,000 square miles." Two of the questions (USEmpls and USChips) suggested to

Table 1

Questions Used in the Experiment, with High and Low Answers.

	Correct Answer	High Answer	Low Answer	X[a]
U.S. Population				
USEmpls: How many U.S. Government civilian employees were there in August 1982?	2,908,025	5,816,050	1,454,013	2.0
USChips: How many pounds of potato chips were consumed in the U.S. in 1982?	972,300,000	2,430,750,000	388,920,000	2.5
Oregon Population				
ORUnemp: How many people received Oregon unemployment benefits in 1982?	188,000	564,000	63,000	3.0
ORTaxes: In 1981 how much money did the state of Oregon collect in all kinds of taxes (including income tax, coporate income tax, gas tax, license fees, etc.)?	$1,608,423,000	$4,021,058,000	$643,369,000	2.5
Speed				
SpHorse: What is the world record speed for a thoroughbred racing horse to race 4 miles?	7 min 10 4/5 sec	10 min 46 1/5 sec	4 min 47 1/5 sec	1.5
SpCar: How many minutes did it take for the 1982 winner of the Datona 500 (a stock-car race) to complete the 500-mile race (rounded to nearest minute)?	195	390	98	2.0
Miscellaneous				
SqMiles: How large is Oregon in square miles?	97,073	291,219	32,358	3.0
Popes: How many Roman Catholic popes have there been?	269	807	90	3.0

Note: The abbreviated titles were not shown to the subjects.

[a]The factor used to arrive at the high and low answers.

us an algorithm using the population of the United States (e.g., "There are about 240 million people in the U.S. If, on average, each person consumes 3 lbs of chips a year..."). Two more questions (ORUnemp and ORTaxes) suggested the use of the population of Oregon in an algorithm. Two questions involved speed (SPHorse and SPCar); the final two questions involved none of the above factors.

The factors chosen to arrive at the "High" and "Low" answers were based on pretest data. In the pretest, subjects were given answers that were two or five times as large as the correct answer or half or one-fifth as large. The subjects were told that the answer was incorrect and were asked whether the answers was too high or too low. From these responses we chose factors to use in the present experiment that we estimated would produce approximately 60% correct reponses. Four of the questions revealed systematic biases in the pretest. For the Popes and USChips questions, more than 50% of the pretest subjects said the answer was too high when the answer given was in fact half the correct answer. For the ORUnemp and USEmpls questions, less than 50% said the answer was too high when in fact is was twice the correct answer.

Instructions. The instructions for the first four questions in the main experiment were the same for all subjects. They emphasized that each answer was wrong and that the subjects' task was to decide whether our answer was too high or too low. Subjects were also instructed to indicate the probability that

their decision was correct. This probability was to be stated as "percent chances," from 50 to 100%. Each question appeared on a single page, followed immediately by the provided answer. The subjects checked "This answer is: ____Too High; ____Too Low." They then expressed their confidence that their response was correct: "The chance that my decision is correct is ____%."

Following these four control questions the subjects read, at their own pace, a detailed tutorial about how to create their own algorithms. Following the tutorial, the final four questions appeared, in the same format as the first four with the additional note, "Show estimates and calculations here," heading the remaining, blank, space on each page.

On the final page of the questionnaire, subjects were asked to estimate the populations of Oregon and the United States.

Twenty-four versions of the questionnaire were prepared. Each version had two high and two low answers in each half. Each version had one U.S., one Oregon, one Speed, and one Miscellaneous question in each half. Across the versions the order of questions varied, with each question appearing equally often in the first and second half and equally often with a high or low answer. The four questions that the pretest had shown were biased were also counterbalanced in the design, with one of the questions for which the pretest subjects believed the answer was larger than it truly was and one for which they thought it was smaller in each half of the quesionnaire.

Tutorials. Each subject received one of two tutorials, the

Forward and the Backward tutorials. The Forward tutorial instructed the subjects to build an estimate of the true answer to the question from facts they already knew or could estimate and to compare their estimate with our answer to see if our answer was too high or too low. The tutorial gave three examples of algorithms. An easy algorithm was shown for the question, "How tall is the Empire State Building?", based on estimates of the number of floors and the height of each floor. Two more complex algorithms were then shown, both for the question, "What was the total attendance at all major league baseball games in 1983?" One of the algorithms was built from estimates of the number of teams, the number of games each team plays per year, and the average attendance per game. The other algorithm, for the same question, was based on the average yearly attendance per team and the number of teams.

The Backward tutorial instructed the subjects to start with our answer and decompose it, using facts they knew or could estimate, to arrive at an implication which they could directly judge as being too high or too low, and to use that judgment to decide whether our answer was too high or too low. The same examples were used, rewritten to conform to the Backward approach. For example, for the height of the Empire State Building, the instructions said,

> Our Answer: 2500 feet.
>
> Here's how to use the method: Start with the number given, 2500 feet, and combine it with a fact you

already know or can estimate. What do you know about buildings? Perhaps you know that in tall buildings each story is approximately 10 feet high. So we use this fact:

2500 feet divided by 10 feet per story equals 250 stories.

Is this believable? Can the Empire State Building have 250 stories? Compare the result of your calculation with your knowledge and your common sense. It just doesn't make sense to suppose that the Empire State Building is 250 stories high. That is much too many stories. So you would conclude that the answer we gave is too high.

The Forwards and Backwards tutorials are given in the Appendix.

Subjects. The subjects were recruited through an ad in the University of Oregon student newspaper and through fliers passed out on campus. They came to a classroom at any time during the specified day, completed the task (and another short, unrelated task) at their own pace, and were paid for their participation.

There were 245 subjects, 136 males and 109 females, with a median age of 23 (range 15 to 58).

Results

Over all subjects, there was only a small increase in accuracy as a result of using algorithms. For the first four questions, 65% of the subjects' decisions were correct; for the

last four, 69% were correct. This increase is marginally significant (2.60 correct answers per subject vs. 2.77 correct answers, $F = 3.58$, $p < .06$). The percentage correct was better in the second half for six of the eight questions; a small decrease in percentage correct was found for USChips (from 65% to 64%) and SpHorse (from 68% to 62%). There was no difference in performance between the group receiving the Forward tutorial (71% correct) and those receiving the Backward tutorial (68% correct).

Scoring algorithms. Each of the 245 subjects was asked to write four algorithms. These 980 algorithms were scored by two coders[1] (with disagreements resolved by discussion) according to the following coding scheme:

F: A forward algorithm, essentially complete.

B: A backward algorithm, essentially complete.

BF: Two complete algorithms, one forward, one backward.

FT: A forward algorithm, complete, with a time-reversal error (see below). Used only for the two Speed questions.

BT: A backward algorithm, complete, with a time-reversal error. Used only for the two Speed questions.

FI: An incomplete algorithm, with enough detail given to recognize it as forward.

BI: An incomplete algorithm, with enough detail given to recognize it as backward.

N: No algorithm or an algorithm so incomplete that the coders could not classify it as forward or backward.

K: The subject did not give an algorithm because the

subject claimed to know the correct answer.

A particular kind of conceptual error was occasionally made with the two speed questions; we called it a time reversal. The essence of such algorithms was, for example, for the SpCar question with the low answer of 98 minutes:

> If a car can go 500 mile in 98 minutes, it goes about 300 miles an hour. That is faster than a race car can go. Therefore, the 98-minute answer is too high.

Here the reasoning is valid until the last statement. If 300 mph is too fast a speed, that logically implies that the answer is too low, not too high.

Similarly, for the SpHorse question with the high answer of 10 minutes 46 1/5 seconds:

> If a horse can run 4 miles in about 10.75 minutes, it is running at a speed of about 22 miles per hour. That is slower than race horses can run, so the answer is too low.

In fact, running too slowly is logically associated with a too-high answer, not a too-low answer.

The algorithm coding scheme did not consider the sensibleness of the algorithm. Some algorithms were quite detailed and sophisticated. Two such algorithms are shown in Table 2. Both were forward algorithms given to the ORUnemp question with a low (63,000) answer. The first algorithm (written by a 21-year-old female) omits the fact that not all

people of working age seek work. In addition, it does not take into account the fact that most people on unemployment do not remain unemployed for an entire calendar year; thus if the unemployment rate were 10% at any one time, many more than 10% would receive benefits in the course of a year. These two omissions created cancelling errors; the resulting estimate of 150,000 was quite close to the true value of 188,000. The second algorithm, written by a 24-year-old male, includes the fact that not all people of working age work or want to work but omits the distinction between the number of people drawing benefits at any one time and the number during the course of the whole year. Thus although his algorithm is quite sophisticated, his answer, 30,000, is low by a factor of more than 6.

Insert Table 2 about here

Most of the complete algorithms used the approaches we expected, such as estimating the length and width of Oregon and multiplying them together to estimate the square miles. A few algorithms used novel approaches. One subject responding to ORTaxes considered state expenditures rather than state income (a not unreasonable approach; Oregon's Constitution forbids deficit spending) by dividing the answer given by the salary of the President of the University of Oregon and assessing whether it was reasonable to suppose that the state employed that many people.

Table 2

Two Algorithms Written for the ORUnemp Question

Population of OR: 2 million

Unemployment in 1982 ≃ 10%

But not all Oregon residents are seeking employment. The baby boomers are in their 30's (i.e., the bulk of the population is middle aged and therefore seeking employment).

If you live 70 years you probably work about 45 years.

45/70 = 64% seeking employment

But I will add 10% for the baby boomers

75% of 2 million = 1,500,000

10% of that = 150,000 out of work

Life expectancy: 76 years

Years eligible for unemployment:

Age 18 to 60 = 42 years

42/76 ≃ 1/2

Population of OR: 2.3 million

1/2 x 2.3 = 1.15 million

% who work or want to work: 30%

30% of 1.15 million ≃ 300,000

10% Unemployment, thus

30,000 unemployed, receiving benefits

Some algorithms showed faulty reasoning, wildly incorrect estimates, or arithmetic errors. These included one subject who calculated the area of Oregon using the formula, Area = (length + width)^2. Another estimated the length of Oregon as 700 miles, noted that there are about 5,500 feet in a mile, and ended with the conclusion, "700 x 55 = 38,500 square miles." Both misestimation and arithmetic error can be seen in this simple algorithm for the area of Oregon: "1000 [miles long] x 3000 [miles wide] = 30,000 square miles." One subject's estimate of the number of Popes involved a listing: One in the US, one in South America, one in Europe.... We did not attempt to score algorithms for their quality of reasoning (except for time reversals), adequacy of estimates, or arithmetic correctness. All the examples given here were coded as complete algorithms.

The subjects were not consistent in following the tutorial instructions to produce Forward (for half the subjects) or Backward (for the other half) algorithms. Counting only the algorithms that could be identified as forward or backward (whether complete or incomplete), 77% of the Forward-instructed group's algorithms were coded as forward; only 51% of the Backward-instructed group's identifiable algorithms were coded as backward. This asymmetry suggests that the subjects found forward algorithms more natural or easier to create. Backward algorithms were rarest for SqMiles--only 13% of all identifiable algorithms--and commonest for USEmpls--59%. The most common algorithm for USEmpls was a backward algorithm along the lines of

the following: If there are 100,000,000 workers in the US, the given answer of 1,454,013 implies that 1.4% of all workers are Federal workers, a percentage that is too low.

Table 3 shows the frequency of different types of algorithms the subjects produced and the corresponding percentages correct. In a high majority of cases, subjects were able to write algorithms that the coders could understand (although some of the algorithms had severe faults). The subjects' decisions as to whether the experimenter-provided answers were too high or too low were more often correct when a complete algorithm was given (70.3%) than when it was not (63.2%). The lowest percentage correct was associated with time reversals (which were made by 13% of the subjects on their Speed question); because this error led subjects to the opposite conclusion, only 15.2% of these decisions were correct.

Insert Table 3 about here

The data shown in Table 3 suggest the hypothesis that subjects who are able to write complete algorithms profit more from the tutorial than those who don't. This hypothesis was tested by dividing subjects into two groups, those who wrote four complete algorithms and those who wrote three or less. (A finer division was not possible because of reduced sample size. For example, only 6 subjects produced no complete algorithms.) For this analysis, the 33 algorithms that contained time reversals

Table 3

Frequency of Types of Algorithms with Percentage Correct, Across all Subjects and Questions in the Second Half of the Experiment.

Type of Algorithm	Frequency	% Correct
Complete		
Forward	506	72.5
Backward	281	71.5
Both	24	83.3
Time Reversal	33	15.2
Subtotal	844	70.3
Incomplete or None	133	63.2
Knows Answer	3	66.7
Total	980	69.3

were categorized as incomplete, which should have the effect of enhancing the probability of finding support for the hypothesis, since these 33 subjects are all scored as having less than 4 complete algorithms and time reversal errors are known to lower radically the percentage of correct decisions. Nevertheless, the data in Table 4 do not support the hypotheses. The main effect of algorithm completeness is highly significant, $F = 13.60$, $p < .001$, and, as previously mentioned, the main effect of the tutorial is marginally significant. However, the interaction that one would expect under the hypothesis does not exist, $F = .02$, $p = 0.88$. Subjects who could write four complete algorithms made more correct decision before the tutorial as well as after it. Thus the most reasonable explanation of the data in Table 3 is that subjects who have the knowledge or intelligence to write a complete algorithm are more likely than others to make a correct decision (with or without the tutorial).

Insert Table 4 about here

Misinformation. Although we could not study the effect of the faulty logic, misestimates, and arithmetic errors that we found in our subjects' algorithms, the experimental design did allow us to explore the effect of two items of information on subjects' performance. At the end of the experiment, subjects were asked to estimate the population of Oregon and of the U.S.

The population of Oregon at the time this experiment was run

Table 4

Mean Number of Correct Decisions (out of 4) as a Function of Number of Complete Algorithms, Before and After the Tutorial.

	Before Tutorial	After Tutorial	
4 Complete Algorithms	2.73	2.92	N = 142
<4 Complete Algorithms	2.42	2.57	N = 103

was about 2,662,000. Our subjects' estimates ranged from 17,000 to 800,000,000, with a median of 2,225,000, first quartile of 1,500,000, and third quartile of 3,050,000. We separated the answers into three groups, high (> 3.5 million; n = 53), low (< 2 million; n = 76), and about right (between these values; n = 115); there was one missing value.

The population of the U.S. was about 235,100,000. Our subjects' estimates ranged from 82,000 to 748,000,000,000 (this estimate may have been a joke, but the next highest estimate was 250 billion), with a median of 248,900,000; first quartile, 210,000,000; third quartile, 600,000,000. We separated the answers into three groups, high (> 300 million; n = 78), low (< 1.75 million; n = 39), and about right (between these values; n = 128). For both the U.S. and the Oregon estimates, the divisions were made at approximately 75% and 133% of the true values.

Each subject received one question for which knowing the population of Oregon would have been helpful (either ORUnempl or ORTaxes) and one question for which knowing the population of the U.S. would have been helpful (either USEmpls or USChips) before receiving the tutorial and one of each type after the tutorial.

Table 5 shows the results for these four questions separated according to whether the question appeared before or after the tutorial, whether the answer provided by the experimenters was too high or too low, and whether the subject's belief about the relevant population was too low, about right, or too high.

Insert Table 5 about here

Here, at last, we find a strong effect of the tutorial. Suppose we give you the high answer to the potato chip question. If you believe that there are fewer people in the U.S. than there are in fact and if use use that false knowledge to evaluate the potato chip question, you should be more likely to decide that our answer is too high than if you had not used your false knowledge. But if we gave you the low answer, using your false knowledge might lead you astray. Because you believe there are so few people in the U.S., you might wrongly conclude that our low answer was too high. Thus, the increase in percentage correct in the first row of Table 5, from 50% to 70%, and the decrease in the fourth row, from 76% to 60%, both suggest that the effect of the tutorial was to cause subjects to use their knowledge about other relevant facts.

A similar argument can be made for those whose population estimates are too high: The tutorial, if effective, should increase performance when the answer is low (strongly supported by the data, 60% to 86%) but decrease performance when the answer is high (weakly supported, 49% to 46%).

This reasoning also predicts a consistent pattern down the columns of Table 5. When the answer is too high, best performance should occur for those who make low population estimates and worst for those who make high population estimates.

Table 5

Percentage Correct for the U.S. and Oregon Questions as a Function of Population Beliefs

	Percent Correct		
	Before Tutorial	After Tutorial	Difference
Answer was High			
Subject's Estimate Low	50	70	+20
Subject's Est. About Right	55	58	+3
Subject's Estimate High	49	46	-3
Answer was Low			
Subject's Estimate Low	76	60	-16
Subject's Est. About Right	70	75	+5
Subject's Estimate High	60	86	+26
Over All Data	62	65	+3

The reverse should be true when the answer is too low. These patterns are seen only in the data collected after the tutorial, suggesting that subjects did not use their knowledge of population size until they were taught to write algorithms.

These results from Table 5 were not subjected to statistical tests because each subject is represented four times in Table 5, unsystematically across the cells (one person, for example, might have too high an estimate for the U.S. population and too low an estimate for the Oregon population).

Calibration. Subjects were asked, for each question, not only to decide whether our answer was too high or too low, but also to assess the probability that their decision was correct. These confidence judgments were analyzed for calibration, that is, the degree to which the assigned confidences matched the correctness of the decisions. The results showed the same overconfidence found in many previous studies (for a review, see Lichtenstein, Fischhoff & Phillips, 1982). For example, for all those decisons assigned a confidence of 90%, only 80% were correct; the percent correct was 84% for decisions assigned a confidence of 100%.

It surprised us to find no difference in calibration between the pre-tutorial and the post-tutorial data. We had supposed that the existence of an aid would increase confidence; since it had little effect on accuracy, this would change the confidence/accuracy relationship. We can only conjecture that the subjects found it so difficult to write their own algorithms

that the existence of this aid did not increase their confidence much.

Discussion

In this experiment we gave subjects questions concerning facts they would be unlikely to know but could estimate and provided subjects with a wrong answer. The subjects' task was to decide whether the given wrong answer was too high or too low. After completing four such items, subjects were given tutorials on how to create algorithms, based on facts they knew or could estimate, to help them in their task. They then completed four more items under instructions to write an algorithm for each one.

These efforts to teach subjects to create their own algorithms was successful in the sense that most subjects were able to write algorithms for the questions we gave them. However, the increase in the accuracy of their decisions as a result of creating and using algorithms, from 65% to 69%, was only marginally significant and unimpressive in size.

Two different approaches to creating algorithms were used. In the Forward approach, subjects were instructed to start with facts they knew or could estimate, from these considerations build an answer to the question asked, and compare their answer with the answer provided by the experimenters. In the Backward approach, subjects were asked to start with the answer provided by the experimenters and, using facts they knew or could estimate, derive some conclusion that they could evaluate from common sense.

There was no significant difference in performance between the Forward-instructed and Backward-instructed groups. The Forward-instructed subjects sometimes created Backward algorithms and the Backward-instructed subjects often created Forward algorithms. Overall, 63% of the codable algorithms were Forward, suggesting that this approach is more natural to the subjects.

The goal of the tutorials on how to write algorithms was to teach our subjects to access their knowledge about related matters and to organize this knowledge coherently. By studying the effects of subjects' beliefs about the population of Oregon (where the experiment was performed) and the population of the U.S., we were able to show that the tutorials were successful in causing the subjects to consider their knowledge about facts relevant to the questions at hand. However, many of the subjects held such faulty beliefs (e.g., that the population of Oregon is 100,000 or 300 million) that using these erroneous facts led the subjects astray almost as often as it helped.

Additional barriers to successful use of algorithms were faulty logic and poor arithmetic skills.

As an aid to decision making, then, this approach is a mixed blessing. Young adults who might be supposed to be above average in intelligence can be taught to access their own knowledge and combine it in logical ways, but their lack of mathematical skills and possession of misinformation hampers their performance.

References

Goodman, S. E. & Hedetneimi, S. T. (1977). Introduction to the design and analysis of algorithms. New York: McGraw Hill.

Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgement under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

MacGregor, D., Lichtenstein, S., & Slovic, P. (in press). Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. Organizational Behavior and Human Decision Processes.

Raiffa, H. (1968). Decision analysis: Introductory lectures on choices under uncertainty. Reading, Massachusetts: Addison-Wesley.

Singer, M. (1971). The vitality of mythical numbers. The Public Interest, Spring, 3-9.

APPENDIX

INSTRUCTIONS TO SUBJECTS

Assessing Quantitative Facts

In this task we will present you with some questions, like "What is the world record time to run a mile?" For each question we provide an answer, like "8 minutes." EVERY ANSWER WE GIVE YOU IS WRONG. Your first task is to decide whether the answer provided is wrong because it is too LARGE a number, that is, too high, or because it is too SMALL a number, that is, too low.

The questions are straightforward; we have not used any "trick" items. We found the answers in various almanacs and the like, and then changed the answers, either by raising them or by lowering them. Some of the answers on your form of the questionnaire are too high; others are too low. In the first part there are only four questions. So please take your time on each one. Think hard about it before deciding whether the answer is too high or too low.

Your second task in this first part is to assess the probability that your decision is correct. Suppose you decided that our answer is too high. Then in this second task we want you to tell us the probability that the answer is, indeed, too high. If you decided that an answer is too low, now we want you to give us the probability that it is, indeed, too low.

This probability is stated in terms of "percent chances." It is a measure of the confidence you have in the correctness of your decision. If you are totally uncertain, so that you could just as well have decided with a flip of a coin, then you have a 50% chance of being right. If you are absolutely certain that you are right, as certain as you are of knowing your own name, then you are 100% certain. A response of 60%, for example, means that there are 60 chances out of 100, or 6 chances out of 10, that you made the right decision. Your answer in this second task should always be a number between 50% and 100%, inclusive. DO NOT USE A NUMBER SMALLER THAN 50 OR GREATER THAN 100.

You can start this task as soon as you are sure that you understand the instructions. Feel free to ask questions.

MORE INSTRUCTIONS

Next, you'll get four more questions, very like the four you just did. Again, we are giving you an answer that is WRONG. Again, we are asking for your judgment: Is the answer we give too high or too low? Again, after you make this decision, you should give us a number between 50 and 100 to express your confidence that your decision is correct.

But this time we want you to use a particular method for evaluating the answer. In a nutshell, this method is to start with the answer we give you and to reduce it, using one or more facts or estimates and some simple arithmetic, in order to arrive at a number you can evaluate more easily, using common sense.

The method will be clearer with a couple of examples. First, a very simple example:

How tall is the Empire State Building in New York City (excluding the TV antenna on top)?

Our answer: 2500 feet

Here's how to use the method: Start with the number given, 2500 feet, and combine it with a fact you already know or can estimate. What do you know about buildings? Perhaps you know that in tall buildings each story is approximately 10 feet high. So use this fact:

2500 feet divided by 10 feet per story equals 250 stories.

Is this believable? Can the Empire State Building have 250 stories? Compare the result of your calculation with your knowledge and your common sense. It just doesn't make sense to suppose that the Empire State Building is 250 stories high. That is much too many stories. So you would conclude that the answer we gave is too high.

Alternatively, you might have approximate knowledge of how many stories there are in the Empire State Building. Let's say that you remember that there are about 100 stories. The the method would go like this:

2500 feet divided by 100 stories equals 25 feet per story.

Is this believable? Even if you have never before thought about how tall one story of a large building is, it flies in the face of common sense to suppose that each story is 25 feet tall. That's more than 4 times as tall as the average person. Again, you conclude that the answer we gave you is too large.

In fact, the Empire State Building is 1250 feet tall and has 102 stories. Thus in fact the average story is $1250 \div 102 = 12.25$ feet. Notice that your estimate of 10 feet per story was a bit off. Nevertheless, the method worked okay, because the answer we gave you was very much off. If you make small errors in your approximations you'll probably still do okay.

Now let's take an example that requires several facts or estimates.

What was the total attendance at all major league baseball games in 1983?

Our answer: 15,186,000

For our first try at this, we'll use three estimated facts:

1. There are about 30 major league teams.
2. Each game is played by 2 teams.
3. Each team plays about 150 games per year.

We start with 15,186,000. Divide this by 150 games per year, getting approximately 100,000, which is an estimate of the number of people attending one or another game on a single day when all teams play. Because there are 30 teams, there are 15 games when all teams play at once. So divide the 100,000 by 15, getting about 6700 attendance at each average game.

Is this number, 6700, sensible? Think of a baseball stadium. They're huge. The average attendance at a single major league game is surely more than this. So the given answer must be too small.

(In fact, there are 26 major league teams and each team plays 162 games per year, but we were close enough in our estimates. The correct answer for the 1983 attendance is 45,557,582.)

There is usually more than one way to approach these questions. For example, suppose I don't have a good idea of how many teams there are, and I don't know how many games they play, but I remember reading that one team had a home attendance, for the year, of less than one million; the article implied that this was very low. So the average attendance for one team must be above one million. I guess it may be 1.5 million (1,500,000). It would have to be at least that high for that article I read to make such a big deal about falling below one million. Using only this one vague estimate, the method goes like this:

15,186,000 total attendance divided by 1,500,000 per team attendance is about 10 teams. Are there only 10 teams? Even if I know very little about baseball, I remember there are two leagues. Only 5 teams per league? I think that's too small. So 15,186,000 is too small, too.

For each of the following four questions and answers, use the method explained above. Start with the answer given and use simple arithmetic and some relevant facts or estimates from your own knowledge to arrive at a number which you can then evaluate using knowledge or common sense.

Please write out enough words and numbers so that we can understand your approach. Try to make it legible and clear (but you don't need to write us a novel).

Take your time. We're giving you only four questions in this part so you can concentrate and do a careful job on each one.

Don't forget to give a confidence rating, too. The rating must be a number from 50 (for complete lack of confidence) to 100 (for utter certainty).

Now that you know this method, please do NOT go back to change any of your previous answers.

After these next four questions, there is one more short page. That completes the experiment. Return the materials and sign for your payment. The experimenter will check to see that you completed everything, that your handwriting is reasonably legible, and that all of your confidence ratings are between 50 and 100 (inclusive). When you finish, take a moment to check these things, too.

You can go ahead as soon as you understand these instructions. Feel free to ask questions.

Thank you for your participation.

MORE INSTRUCTIONS

Next, you'll get four more questions, very like the four you just did. Again, we are giving you an answer that is WRONG. Again, we are asking for your judgment: Is the answer we give too high or too low? Again, after you make this decision, you should give us a number between 50 and 100 to express your confidence that your decision is correct.

But this time we want you to use a particular method for evaluating the answer. In a nutshell, you will use your own knowledge of related facts to construct your own answer, then compare your answer with ours.

The method will be clearer with a couple of examples. First, a very simple example:

How tall is the Empire State Building in New York City (excluding the TV antenna on top)?

Our answer: 2500 feet

Here's how to use the method: Forget our answer for a moment, and construct your own. What do you know that's relevant? Perhaps you can estimate the number of stories in the Empire State Building. Say, about 100. And you can give a reasonable estimate of the height of an average story, say about 10 feet. These two facts or estimates drawn from your own knowledge can be put together to get an estimate of the target quantity:

100 stories times 10 feet per story equals 1000 feet, height of building.

Your estimate, 1000 feet, is much lower than our answer. So our answer must be too high.

That's all there is to it. Search your memory and use your common sense to get facts or estimates that are relevant. Put these numbers together using simple arithmetic to arrive at your own estimate. Compare your estimate with our answer.

In fact, the Empire State Building is 1250 feet tall and has 102 stories. Thus in fact the average story is 1250 - 102 = 12.25 feet. Notice that your estimate of 10 feet per story was a bit off. Nevertheless, the method worked okay, because the answer we gave you was very much off. If you make small errors in your approximations you'll probably still do okay.

Now let's take an example that requires <u>several</u> facts or estimates.

What was the total attendance at all major league baseball games in 1983?

Our answer: 15,186,000

For our first try at this, we'll use four estimated facts:

1. There are about 30 major league teams.
2. Each team plays about 150 games per year.
3. Each game is played by 2 teams.
4. The average attendance at any one game is about 15,000.

Put these all together. Thirty teams times 150 games is 4500 team appearances per year. But since each game requires two teams, 4500 ÷ 2 = 2250 total games per year. Games per year times average attendance per game equals total attendance:

2250 times 15,000 = 33,750,000

Our new estimate is 33,750,000. That is much larger than the answer provided (15,186,000), so we conclude that the answer provided is too low.

(In fact, there are 26 major league teams and each team plays 162 games per year, and the average attendance per game is 21,632, but we were close enough in our estimates. The correct answer for the 1983 attendance is 45,557,582.)

There is usually more than one way to approach these questions. For example, suppose I don't have a good idea of how many teams there are, and I don't know how many games they play, but I remember reading that one team had a home attendance, for the year, of less than one million; the article implied that this was very low. So the average attendance for one team must be above one million. I guess it may be 1.5 million (1,500,000). It would have to be at least that high for that article I read to make such a big deal about falling below one million.

But how many teams are there? I remember there are two leagues. Each league must have at least 10 teams, for a total of 20 teams. I don't think they each have as many as 20 teams, for a total of 40 teams. Let's try an estimate a bit below the middle of that range, say 25 teams. Twenty five times 1.5 million attendance for each team gives a total attendance of 37,500,000. That is much higher than the provided answer. Even if I had used my low guess for the number of teams (20), I still would have gotten an estimate larger than the one provided. So it looks like the answer given is too small.

For each of the following four questions and answers, use the method explained above. Use simple arithmetic, some relevant facts or estimates from your own knowledge, and common sense to arrive at your own estimate of the answer. Then compare your answer with the one given.

Please write out enough words and numbers so that we can understand your approach. Try to make it legible and clear (but you don't need to write us a novel).

Take your time. We're giving you only four questions in this part so you can concentrate and do a careful job on each one.

Don't forget to give a confidence rating, too. The rating must be a number from 50 (for complete lack of confidence) to 100 (for utter certainty).

Now that you know this method, please do NOT go back to change any of your previous answers.

After these next four questions, there is one more short page. That completes the experiment. Return the materials and sign for your payment. The experimenter will check to see that you completed everything, that your handwriting is reasonably legible, and that all of your confidence ratings are between 50 and 100 (inclusive). When you finish, take a moment to check these things, too.

You can start as soon as you understand these instructions. Feel free to ask questions.

Thank you for your participation.

ONE LAST, SHORT PAGE

Please make estimates of the following as carefully as you can:

What is the population of the United States? _____

What is the population of Oregon? _____

That's the end. Please review to make sure you didn't leave anything out. Then return this questionnaire for your payment. We would be very grateful if, for the next week, you would NOT discuss this experiment with anyone who has not participated in it.